**Course:** SMU Data Science - 2024
**Project 2, Group 1** - Crowdfunding Report
**Professor:** Alexander Booth
**Contributors:** S. Hawkins, J. Cisneros, and R. Yusuff

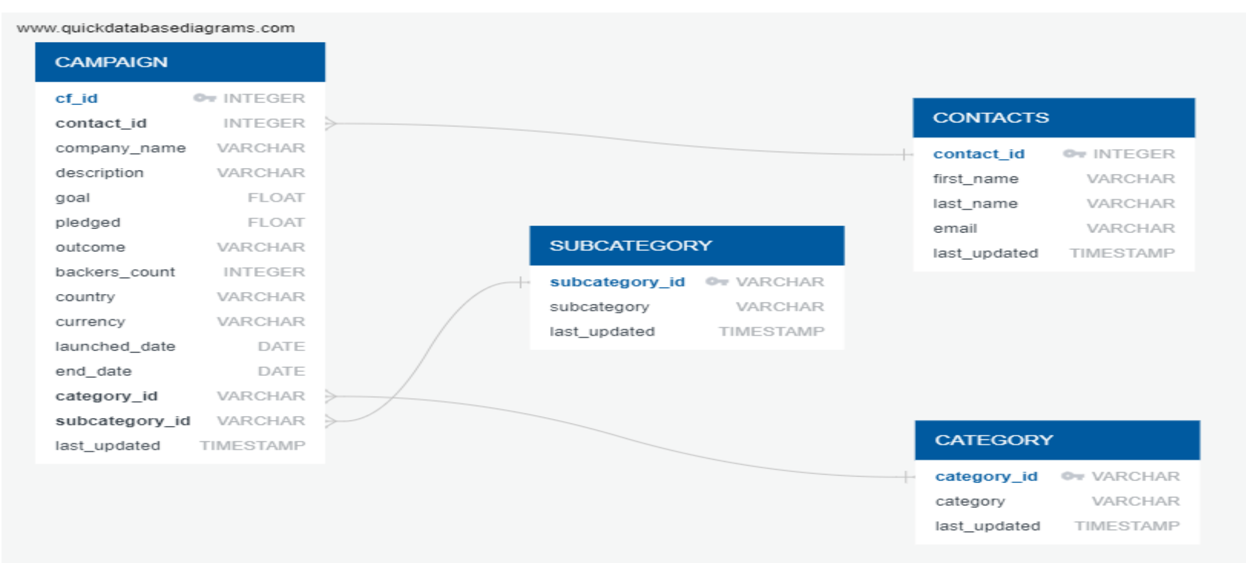
# DATA SCIENCE REPORT: A STATISTICAL ANALYSIS OF CROWDFUNDING CAMPAIGNS

To date, Crowdfunding Campaigns have experienced a consistent, upward trend for more than two decades. These campaigns have served as a distinct and innovative strategy, linking investors and creators who are seeking financial backing for their business venture or special project across diverse industries. Using specific tools and techniques, data scientists can contribute to the efficiency, accuracy, and scalability of crowdfunding campaigns by taking specific data retrieved from its reputable sources, implementing the ETL process (extract, transform, and load), and making it suitable for statistical analytics and reporting.

For this project, we built an ETL pipeline by reading the original Crowdfunding and Contacts Excel files into a Pandas Data Frame in Python. Next, we performed data cleaning and extraction methods such as dropping, reordering, and renaming columns, converting rows in a data frame into a dictionary, checking the accuracy of the data types, and creating four new data frames exported into CSV files. Next, we sketched an Entity Relationship Diagram (ERD) to create table schemas using our cleaned CSV files and exported those into a Postgres Database. Finally, we queried the data using SQL language syntax and functions, garnering insights to support better decision-making for investors while enhancing the overall data integration, quality, and management processes for the crowdfunding campaign platforms.

**Fig. 1 - Crowdfunding Campaign ERD**

The following is a visual representation of our completed ERD (Fig. 1). Through the process of data modeling, we imported our CSV data frames to create four table schemas: *Category*, *Campaign*, *Contacts*, and *Subcategory*. Then, we connected each table entity as appropriated by the relationship between its similar and connecting attributes, if any. For example, the *Campaign* and *Contacts* tables were connected by the *Contact_id* attribute.



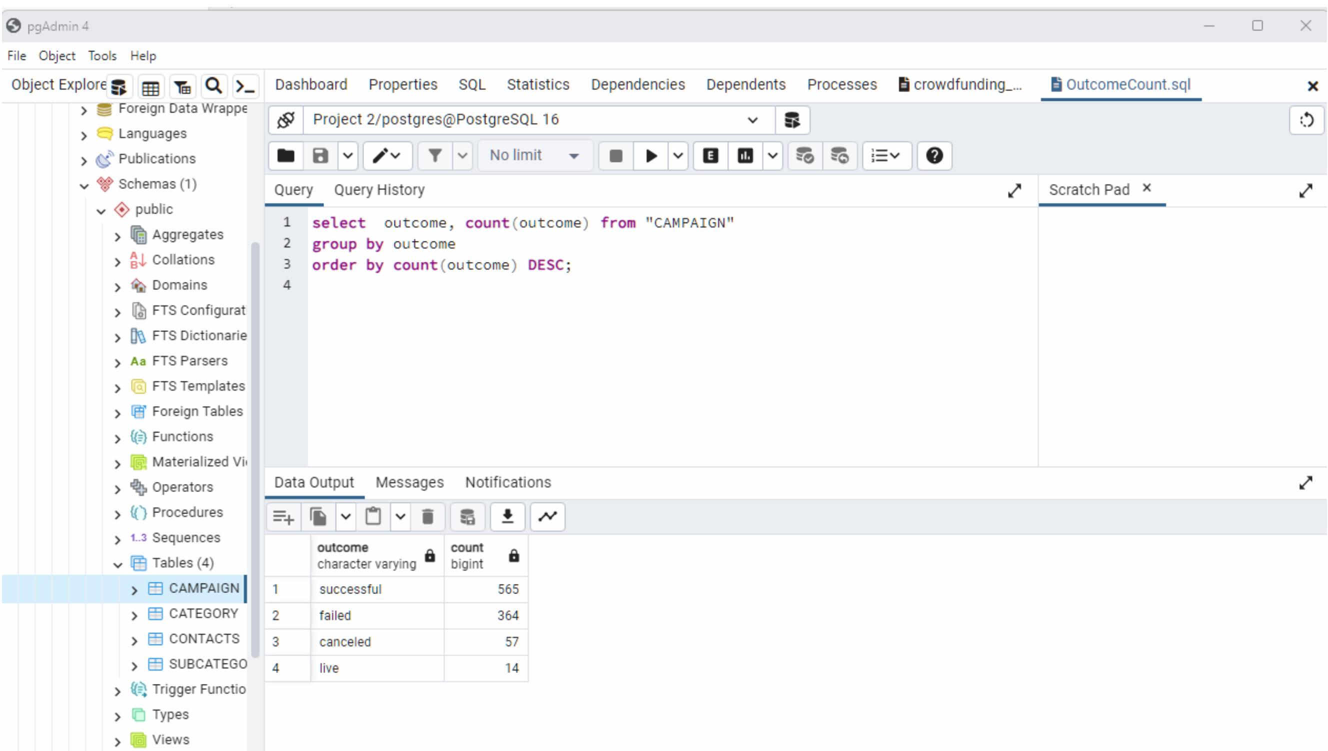
**Figs. 2 - 4: Database, Querying, and Results**

Using PostgreSQL, we created a new database titled “Crowdfunding Database.” Then we created our four tables to run five SQL queries as shown below in Figs 2a-2c, 3, and 4.

**Fig. 2a: Query 1a**
**Fig. 2b: Query 1b**
**Fig. 2c: Query 1c**

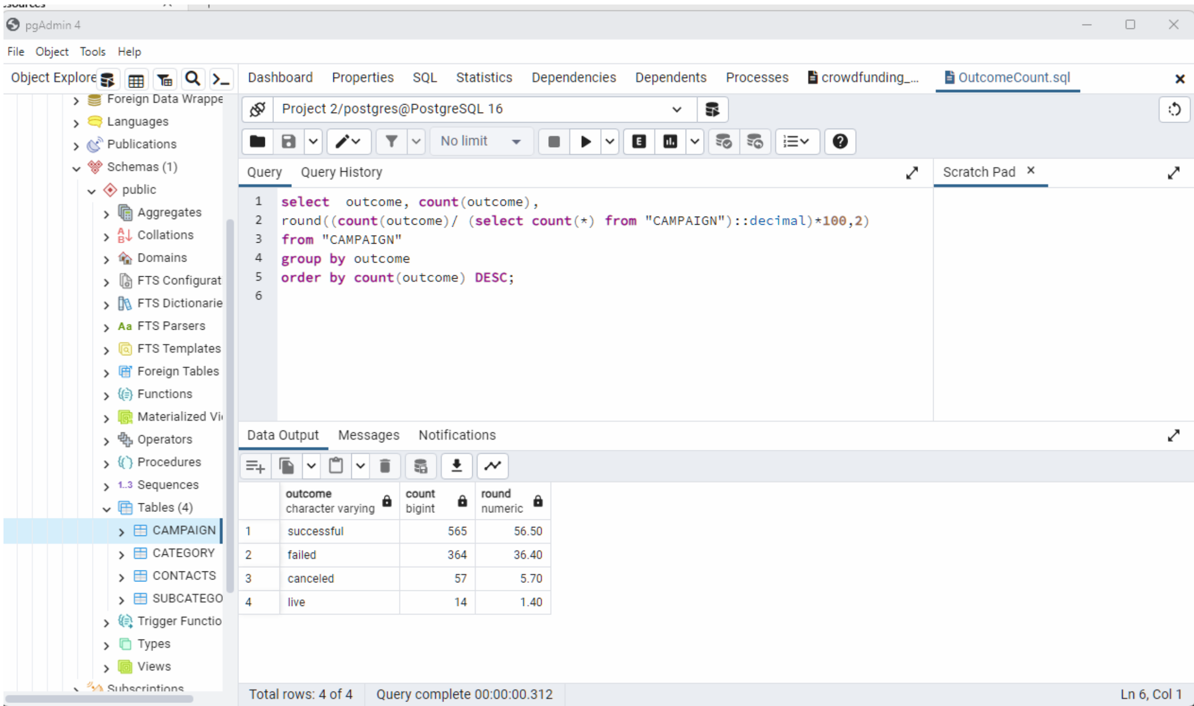
Using the *Select*, *Group By*, *and Order By* clauses, from the ‘CAMPAIGN’ table, we sorted the outcomes into four categories-successful, failed, live, and cancelled-and calculated the total occurrences of each from 1,000 campaigns to determine the outcome counts by descending order from highest to lowest (Fig. 2a).

FIG.2a: Query 1a




Next, we created a fourth query where a new column titled “round” was generated. In this column, we took the total outcome counts within each of the four categories and divided by the total outcome counts from the ‘CAMPAIGN’ table. Then we identified the percentage values of each from the total counts of each outcome type. The four outcomes were arranged in descending order from the highest to lowest percentage. This ensured the output data from Query 4 (Fig. 2b) returned accurate values counts and percentages that coincided with the output data from Query 1 (Fig. 2a). We used this clause to provide proof that all the data needed to conduct our analysis was imported correctly following the ETL process.

FIG.2b: Query 1b




Using the count of outcomes from the "CAMPAIGN" table, we combined multiple select statements implemented the "union all" function to create a new result that displayed the count of each outcome combined into one total count. Having the combined total count of outcomes provides a system of checks and balances for writing and solving equations from datasets with large amounts of numerical data. Knowing the total count of outcomes equaled to 1,000, helped ensure accuracy and reliability of each individual category outcome count and each percentage value. This adds a layer of confidence that our final conclusions are valid, maintaining the integrity of our findings, insights, and future implications.

FIG.2c: Query 1c

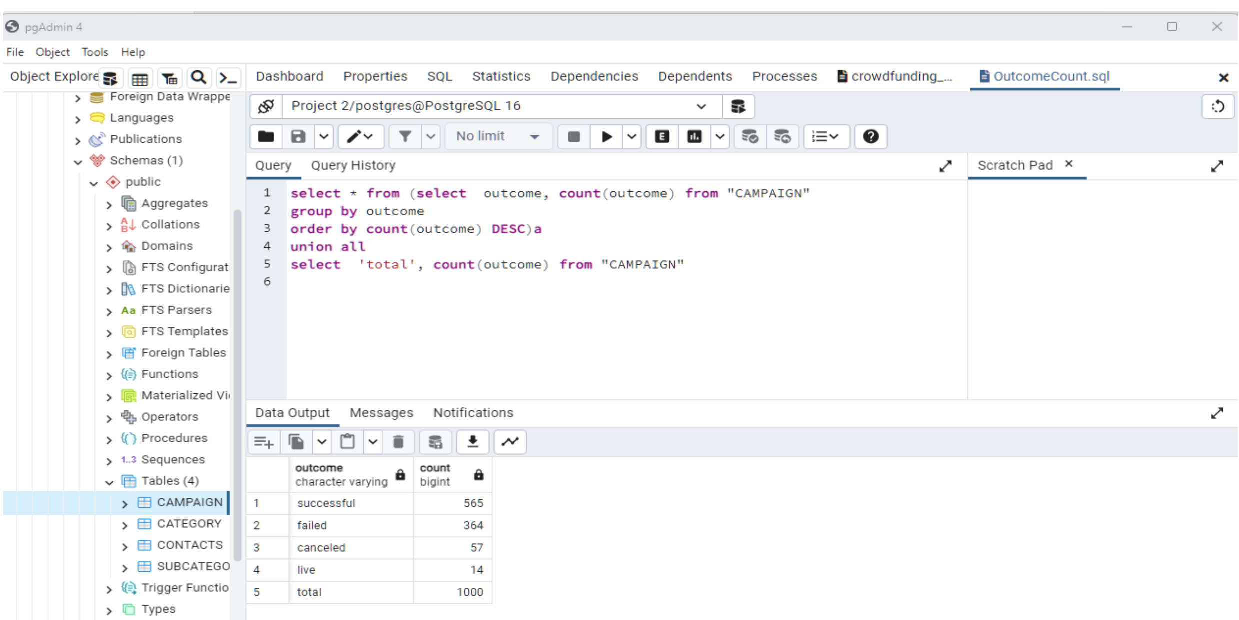


**Fig 3: Query 2**

Using the JOIN query on the *contact_id* attribute that connected the *Campaign* and *Contacts* tables, we were able to retrieve the *cf_id*, *first_name*, and *last_name* for each *contact_id* by joining these tables based on the relationship between a primary key and a foreign key (Fig. 3). From there, we ran an order by clause to arrange the columns in descending order by last name, sorting the information in an organized manner. With this query, a campaign owner and his contact information can be quickly located, as well as, easily accessible whenever necessary.

FIG. 3

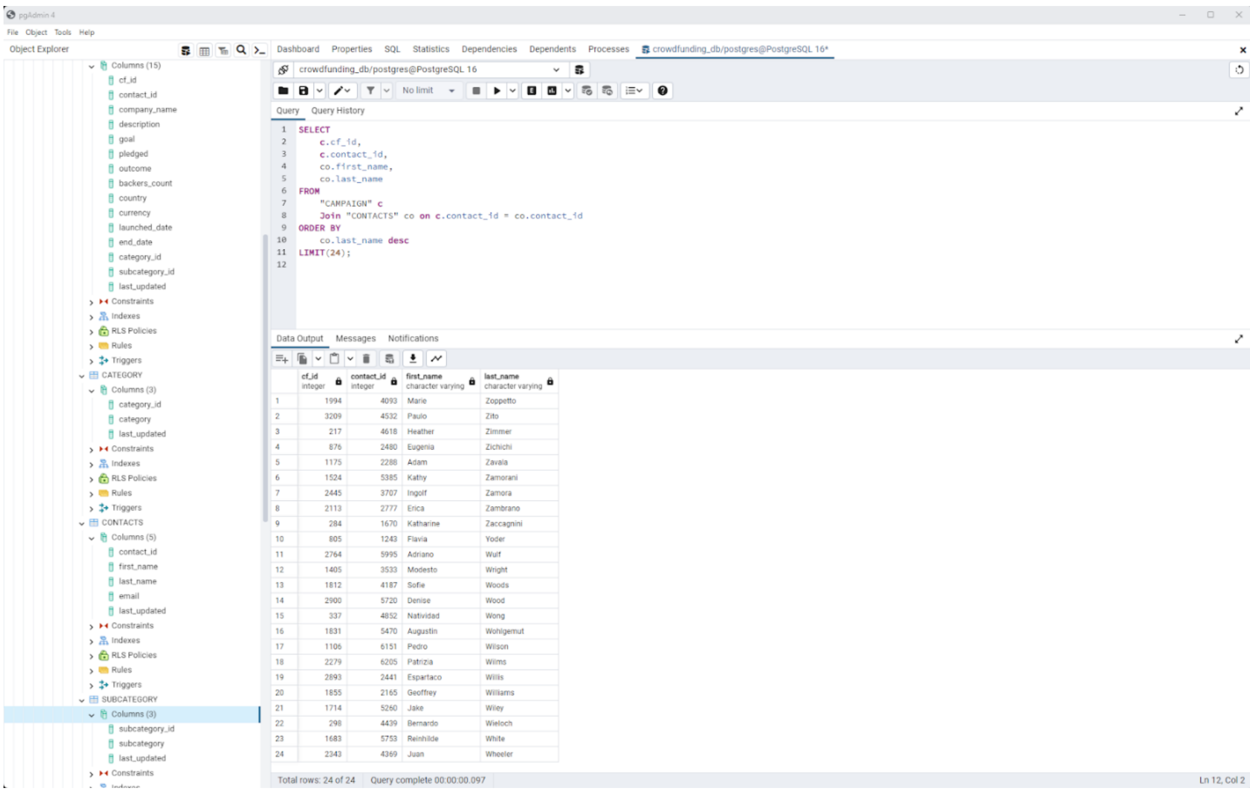
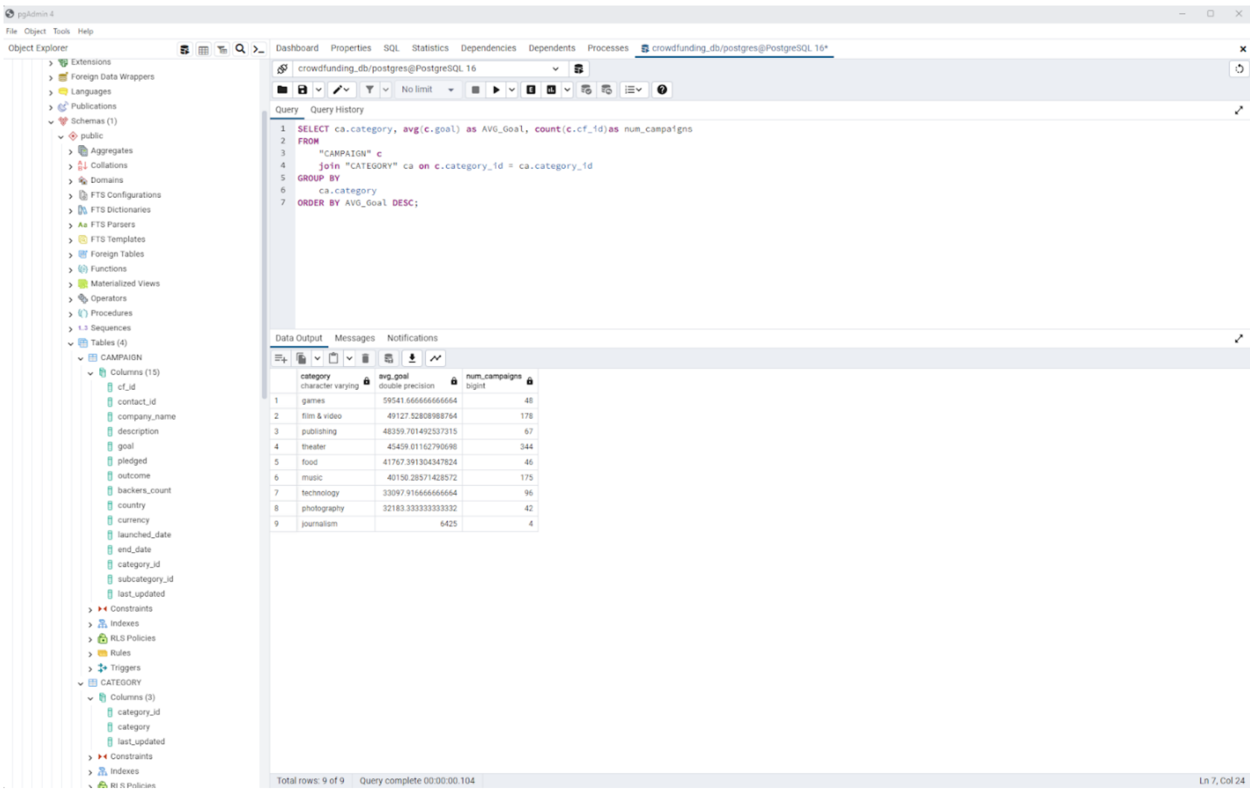



**Fig 4: Query 3**

Another query we wrote was an aggregation query for a given use case where a user would need to locate which category has the highest campaign goals and how many campaigns are present in their respective category. We accomplished this task by locating each of the categories we offer, a total of nine, as well as the average goal for that category, and the number of campaigns in the category. We located said information by joining two tables based on a common key, *category_id* found within the tables, *Campaign* and *Category*, then the average function was performed on the *goal* column, which calculated the average goal of our campaigns. Next, we performed a count function of the cf_id column renamed as *num_campaigns*. Finally, we grouped this information by *category* and ordered by *avg_goal* in descending order (Fig. 4). By performing this query, we were able to determine the category with the highest and lowest average goal amount, which was Games and Journalism, respectively. One noteworthy insight that was similar between the Games and Journalism categories was that they both had a small number of total campaigns when compared to a few of the other seven categories within the dataset. With the specific clauses and functions used to create Query #3, crowdfunding investors and stakeholders can determine that since Games require a larger budget in comparison to Journalism, more time, money, and resource allocation would be needed to fulfill the goals and objectives for a Games project or business venture to be classified as successful in the end.

FIG. 4




## Conclusions, Future Work, and Implications

Based on the final analysis of our completed ETL processes and our five queries, in the future, a more insightful analysis could include queries that factor in which Company's received the highest percentages in pledged amounts compared to their goal amount and determine which company's campaign would benefit most from additional marketing or promotional efforts. Finally, we could perform a query of the country with the total count of successful and total count of failed campaigns to determine which categories fared better based on country, which would spark an analysis that drives insights and better decision making on how to increase the chances of success for categories that previous data shows are typically expected to fail in each country.